検索結果の絞り込みのために有用な語集合の特定

手島 亮太† 岡部 正幸‡ 梅村 恭司†

†豊橋技術科学大学 情報・知能工学系
‡豊橋技術科学大学 情報メディア基盤センター
†teshima@ss.cs.tut.ac.jp, okabe@imc.tut.ac.jp umemura@tut.jp

1 はじめに

ある話題について検索した結果として，一度に読み切れない数の文書が出力された場合，文書の数を絞り込むために追加の語を検索語に加えることが良く行われているけれども，このときに効果的な語を考えることは，話題の内容に関する知識が必要な作業である。この知識がない場合は，検索結果の一部分を読み，次の語を考えることになる。本来ならば，検索結果となった文書集合の全体を俯瞰して，効果的な語を特定すれば良いが，そのためにはそれぞれの文書を読み解く必要があり，全体を俯瞰するのは人間にはコストの高い作業である。

このコストを軽減するために，検索結果のなかから，絞り込みに効果のある語の集合をコンピュータが特定するというタスクを考えた。この語の集合を検索結果から特定すれば，検索の絞り込みのときに役に立つ。このタスクのためには，下記の機能が必要と考えた。(条件 1) 少なくとも，単独でも検索に効果がある語を特定すること。(条件 2)検索を継続することを想定し，語を追加しても候補が十分に残る語を特定すること。(条件 3) 語の集合の大きさを無駄に大きくしないために，出現の相関の大きい２つの語があった場合はどちらか一方を選ぶこと。これらに加えて，下記の機能を想定した。すなわち，(条件 4)検索対象のなかに未知語が多く含まれる事を想定し，語の特定に形態素解析プログラムに代表される対象分野に依存した情報を含むものを使わないで動作すること。

本稿では，以上の機能を備えたシステムのプロトタイプの構造を述べ，プロトタイプのために利用したシステムと，その問題点，および，その改善方法を述べたあと，システムの具体的な動作例を示す。

2 アプローチ

システム動作を確認するには，NTCIR-1 のテストコレクション[1]を利用して作成した擬似的な検索結果集合を作った。これは，論文の抄録の情報検索のコレクションである。このなかから，特定の話題を表現する文字列を指定し，その文字列を含む文書集合を入力文書とした。表１に，入力文書集合と，対象となる文書数を示す。いずれも，すべてを読むには量が多いものとなるように話題を設定した。

表 1, 利用した文書集合

文書集合	文書件数	共通する話題
ALGO	9477	アルゴリズム
CLUSTER	470	クラスタリング
COMPILER	588	コンパイラ
COMPRESS	7132	圧縮
NOISE	3641	雑音
OPT	4081	最適化
PC	646	パーソナルコンピュータ
ROBOT	2712	ロボット
SOFT	4092	ソフトウェア

対象分野の情報を使わない条件（条件４）は有用だが独特のものである。形態素解析を使わない条件で検索に効果のあるキーワードを使う方法[4]があり，この方法の出力結果の中から，目的の語の集合を選び出すアプローチを採用した。ただし，文献[4]の方法では，検索に役立つ文字列を出力するものの，語とはいえないものも出力されるので，検出後の後処理として，語としてふさわしいものを選ぶことを行う。

その後，Zamir らの Suffix Tree Clustering[2]と類似な処理，統計的なヒューリスティクスによる選別と重複をとる操作を行うことで，等価的に検索結果のクラスタのラベルに相当するものを特定し，そのリストを出力することにした。

これらを実装するときには，任意文字列長による統計値を求める umemura らのアルゴリズム[5]をベースにし，さらに任意文字列間の重複出現文書数を扱えるような拡張を行ったものを使った。

3　抽出キーワードの改善

武田の処理は Church によって提唱された反復度(adaptation)[3]という統計量を用いる。これは重要な単語は繰り返すということを着眼点として，形態素解析システムを利用しないキーワード抽出を実現しており、この反復度が文章を特徴付ける語の抽出に有用であることが述べられている。ここで、文字列 x が一回以上出現する文書数を df(x)、二回以上出現する文書数を $df_2(x)$ とする時、反復度 adaptation は次式で表される。

$$adaptation = \frac{df_2(x)}{df(x)}$$

武田らの反復度によるキーワード抽出は、テキスト上のすべての分割を考え、それによって作られる任意の部分文字列を対象としていることから抽出語に特有の誤りを含むことがある。その誤りと対策について順に説明を行う。

3.1 区切り文字の改善

前述のキーワード抽出には「、」や「。」などの句読点、「の」や「は」などの助詞がキーワードの先頭・末尾に残るという問題がある。これは、キーワードである語句に特定の助詞や句読点が繰り返し利用されることで、助詞や句読点を併せたものがキーワードとして抽出される問題である。本研究では、そのような誤りであるキーワードを減らすために、句読点や助詞、更に自立語を区切るようにして使われる記号などの文字列に対し、それらの文字列がどのような統計的性質を持つか調査を行い、文字列が持つ統計量から句読点や助詞などを区別なく特定した。以後、このようにして特定した語句を区切り文字と呼ぶこととする。

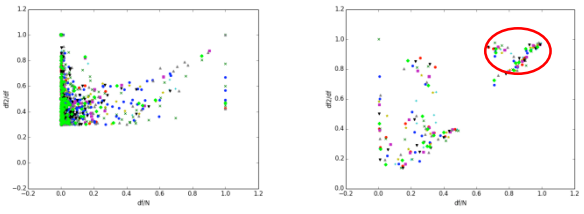

図 1, キーワード(左) と 句読点・助詞・記号(右)の反復度の分布

表 2, 統計量から求められる区切り文字の一例

、(読点) / 。(句点) / ―(長音符号) /
い / し / す / た / て / で / と / に / の /
は / る / を /

3.2 東京都-京都問題への対処

東京都-京都問題は、形態素解析システムをあえて使わずに，文字列の統計値で処理を行うために生じる問題である。具体的には，ある語が他の異なる語の部分文字列であると統計値が重なってしまい，正しく処理されないという問題である。例えば、「東京都」は「東京」という地名と「都」という行政区分の複合語であり、「京都」は「京都」という地名である。従って、「東京都」と「京都」の二語は全く異なるものであるが、どちらも文字列上では「京都」という文字列の一致が見られる。そのため、統計量を求める際に二つの語を意味解釈などで区別しなければ、本質が異なるものであるにも関わらず「京都」の頻度統計量に「東京都」のものが加わるという問題がある。

今回の目的では，出力する語は，比較的に分野特定能力があり，長い文字であることが期待できる。そのため，東京都-京都問題の対策として、「京都」のように他の語に内包される語を結果の候補から外すこととした。

3.3 文章語句の影響の回避

キーワード抽出には反復度 df_2/df を用いているため、出現文書数が少ない語句では偶然の一致や説明の繰り返しなどで「○の□における△」のような幾つかの語から構成される語句(以下、文章語句)がキーワードとして抽出されることがある。本稿では先述したように他のキーワードに内包されるキーワードを最終結果から削除する処理をするので、文章語句がキーワードとして誤り抽出されると，それに含まれる語が候補から削除されるという問題がある。これを回避するため，このような文章語句は区切り文字を内部に含むものだと仮定し、df に対してどのように出現するかを調査した。ただし、長音符号などの解析に不適切な一部の語は、調査に用いる区切り文字としないこととした。

図 2 は、複数の文書集合に対して調査を行った結果である。ここで縦軸の ratio は、

横軸の df を df_Min とすると次式から求められる。

$$ratio = \frac{df \geq df\_Min\text{を満たす文章語句の総数}}{\text{文章語句の総数}}$$

本来は，システムがこの分布を解析し，どのような頻度を打ち切りとするかの指針が必要であるが，今回は複数の入力データに対しての分布が似ていることと，文章語句の 9 割削減を目的として df = 4 を閾値に設定して，文章語句かどうかを判定することにした。ある文字列が文章語句と判定すると，これはキーワードとは考えず，その文字列に含まれる文字列も結果に残ることになる。

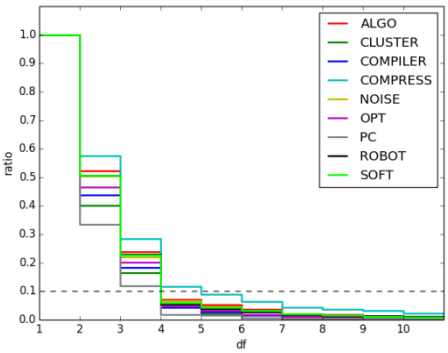

図 2, 区切り文字を含む語句の df に対する出現割合

4 語集合の特定方法

抽出キーワードから目的の語集合に含まれるものを選ぶときには，条件 2 より数多くの文書で取り上げられることが考えられる。一方で条件 3 はそれと相反する性質がある。条件 3 の例は，「ヒト型ロボット」は「二足歩行ロボット」の分類の一つであるが、このような分類を示す語句は共に用いられることが考えられる。このような場合、文書集合の話題を知るには一方が得られれば検索の絞り込み候補があることになり、出力のリストを肥大させないためにもどちらかを選ぶべきである。

本稿では、条件 2 と条件 3 を両立する方法の一つして，表 3 の方法を採用した。ここで述べる「カバーしている範囲」とは，結果に含まれる語を含んでいる範囲であり，カバーしていない範囲とは，現在の語集では追加の検索，候補とならない文書と考えられる。ここで，最初に，最もカバーするキーワードを選んだ瞬間に，カバーしていない範囲がなくなるようにも思えるが，包含関係による選別をしているので，そのようなキーワードは排除され，結果としてそのようなケースはほぼないことがわかっている。ここで、表 4 は文書集合の重要語句を 100 個特定したときに提案手法がどのような結果を与えるかを示したものである。このとき、被排除キーワードとは，語の出現頻度であれば出力結果に含まれるが、この手続きの結果で，選ばれないようなキーワードのことである。また、最も重複するキーワードとは，対応する被排除キーワードと重複出現する文書数が最も大きいようなキーワードのことである。表 4 にあるように「アクチュエータ」の分類に類する「空気圧駆動」や、「学習制御系」と意味が近い「学習制御法」など、分類関係や意味が近いものなどを同時に選ばないようにしている結果が得られたため，条件 2 と条件 3 を両立した手順であると考えられる。

表 3, 文書集合の語集合の選定手順

1. 文書集合全体をカバーされていない範囲とする。
2. まだカバーされていない範囲を最もカバーするキーワードを選択する
3. カバーされていない範囲から,選択されたものでカバーされたものを除く
4. 与えられた個数の語数が選ばれるまで 2 へ行く。

表 4, 語の選定の一例

被排除キーワード	最も重複するキーワード
空気圧駆動	アクチュエータ
歩行パターン	二足歩行ロボット
サブシステム	自律分散システム
軌道生成	二足歩行ロボット
微分方程式	二足歩行ロボット
パラメータ変動	アクチュエータ
学習制御系	学習制御法
モーション	アクチュエータ
最適制御	評価関数

5 文書集合の語集合の特定結果

表 1 の文書集合 ROBOT における文書集合の重要語句を 100 個特定した結果が表 5 になる。ここで、正しい抽出語とは人間が

正しく読み取ることが出来る語であり、誤った抽出語はそうでない語を示している。

ここで，検索の絞り込みの能力についての評価になっていないのは，語の選別のアルゴリズムにより，想定した絞り込みの能力があることが保証されているからである。むしろ，ユーザが利用するときに，出力をみて，それが判別できるかどうかを調べるのが実際の状況に近いと考えたからである。

表 5, 文書集合 ROBOT から抽出された語

正しい抽出語 (93[%] = 93/100)	誤った抽出語 (7[%] = 7/100)
アクチュエータ　ロボットハンド	・マニピュレータ
制御方式　遺伝的アルゴリズム	道追従制御
モジュール　ナビゲーション	力センサ
二足歩行ロボット　協調作業	きロボット
遠隔操作　対象物体	関節マニピュレータ
プロセス　フレキシブルアーム	マニピュレータの逆
評価関数　溶接ロボット	移動ロボット間
演算遅れ時間　組立作業	定性的な
超音波センサ　外乱オブザーバ	
姿勢制御　インタフェース	

表 5 の結果は特定した語句が直ちに文書集合の話題を特定できることには繋がらないが、少なくとも統計量を用いた処理だけで意味のある語が取れることを示している。また、誤った抽出語にある「・マニピュレータ」の中点が残る問題や「道追従制御」(正しくは、「軌道追従制御」)の語の一部が欠ける問題は、武田らのキーワード抽出での誤りの訂正が十分でないケースであると考えられる。

この表において「正しい抽出語」は，「人間にとって意味のある語である」という意味であり，検索の絞り込みという観点からの評価でないけれども，「ロボット」という検索結果の絞り込みの候補と，考える有用そうなリストがとれたと解釈できる。具体的には，特定した文書集合の重要語句からはロボットの種類を示す「二足歩行ロボット」や「溶接ロボット」、ロボットの動作を示す「協調作業」や「遠隔操作」など、ロボットが持つ話題として，検索の絞り込みのために自然な語である。

6　まとめ

本稿では、ある特定の検索対象を利用して，検索結果から辞書を用いないで，絞り込みのための語が取り出せることが実際にできることを示した。

絞り込みのためという概念を，条件１：検索する価値があるという条件，条件２：絞り込みすぎないという条件，条件３：冗長な語が同時に含まれないという条件，さらに，対象に未知語であって，重要な語があるはずである前提で，条件４：形態素解析などの対象に依存した辞書情報はないという条件に整理し，実際にそれを実装した。

今後の課題としては，このシステムの動作は，想定したテストデータに特化して閾値やヒューリスティクスを追加しているので，それが別の検索課題でも正常に動作するか確認する必要がある。

また，現在の実装は，選別にかかるまでの時間の評価を行っていない。現状，計算量的に，平均速度として O(N log(N)) （ただし, N は入力の検索結果の文字列の長さ）になるように想定しているが，無駄な処理や最悪ケースにおいて，計算のオーダーが悪化する可能性が残っており，そこを再検討して実装し，評価する必要がある。

参考文献

[1] Noriko Kando et al. NTCIR: NACSIS Test Collection Project. 20th Annual Colloquium of BCSIRSG, pp.25-27,(1997)

[2] Oren Zamir, Oren Etzioni. Web Document Clustering: A Feasiblity Document, 20th ACM SIGIR Conference, pp.46-54(1998)

[3] Kenneth W. Church, Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to *p/2* than p^2, *Coling*, pp.173-179 (2000).

[4] 武田善行, 梅村恭司:キーワード抽出を実現する文書頻度分析, *計量国語学*, Vol.23, No.2, pp.65-90 (2001).

[5] Kyoji Umemura, Kenneth W. Church, Substring Statistics, CICLing '09, pp.53-71, (2009)